



Universidad
Zaragoza

Trabajo Fin de Grado

Diagnóstico automático de patologías en la
región lumbar

Autor

Ricardo Teruel Arauzo

Director/es

Manuel Doblaré Castellano

EInA

2019

Resumen

Este proyecto pretende crear un programa que sea efectivo diagnosticando patologías de la región lumbar de la columna vertebral. Los datos con los que se contaban consistían en las coordenadas cartesianas de 66 puntos correspondientes a la delineación de las cinco vértebras lumbares y el coxis (8 puntos cada una) y 18 puntos que representaban tres capas de tejido blando (6 puntos cada capa) de 221 pacientes. Mediante el uso de algoritmos de aprendizaje automático, y con la aplicación del método deducido al final del proyecto se han conseguido resultados oscilantes entre el 75-85% de eficacia con un mínimo de 80% y de hasta un 100% de detección de pacientes enfermos, que probó ser el mayor desafío, ya que, en un principio, los mejores algoritmos tal y como están disponibles, para la misma eficacia sólo alcanzaban detecciones del 40%.

El proyecto y todo lo que engloba es aplicado al problema concreto de las patologías de columna vertebral en la región lumbar. Sin embargo, pueden extrapolarse a otros campos donde también sea necesaria la clasificación de elementos en un conjunto, ya que hace uso de un método original que fue desarrollado para enfrentarse a la problemática impuesta por la poca representación de la clase de interés. Este método puede considerarse una nueva aproximación al *Bootstrapping* (1) o al *Boosting* (2).

Contenido

Resumen.....	1
Introducción	3
Patologías	3
Problemática	4
Objetivo.....	5
Material y Métodos.....	6
Datos Disponibles.....	6
Técnicas de clasificación.....	7
Algoritmos de Aprendizaje Automático	7
Criterios de Evaluación.....	7
Comparación de Algoritmos.....	9
Purga de lista de entrenamiento.....	9
Algoritmos en Tándem	10
Scores SVM.....	10
Generación Tabla Probabilidad	10
Curvas ROC	11
Separación grupo de verificación.....	12
Resultados.....	12
Primeras Tablas	12
Mejoras	14
Tabla probabilidades	16
Repetición de las tablas iniciales.....	17
Curvas ROC	18
Verificación.....	20
Conclusiones	22
Futuro	23
Bibliografía	24

Introducción

Las lesiones de espalda son comunes hoy en día, las hay causadas por la degeneración natural o bien por eventos puntuales como golpes o accidentes. En algunos casos la degeneración se ve acelerada por tareas repetitivas que perjudican a la columna, como puede darse en trabajadores o deportistas, o por malas posturas, las cuales se acentúan en un modo de vida sedentario. El diagnóstico de estas condiciones es llevado a cabo por profesionales de la salud con mucha experiencia y que han de invertir mucho tiempo en cada diagnóstico. Estos esfuerzos se multiplican por la frecuencia de dichas lesiones.

Afortunadamente el aprendizaje automático está en auge, y, entre la gran cantidad de campos en los cuales puede aplicarse, está la clasificación automática de elementos en un conjunto. Estas técnicas pueden aplicarse a la detección de piezas defectuosas en una cadena de montaje, búsqueda de imágenes que responden a ciertos patrones, o, para el caso que nos atañe, clasificación de pacientes que sufren o no ciertas patologías en función de la imagen proporcionada por la RMN (Resonancia Magnética Nuclear).

Con el desarrollo de un programa suficientemente preciso, los médicos podrían contar con una herramienta que les ayudase a decidir en el diagnóstico para poder ejercer su profesión de manera más eficiente, pudiendo aumentar su productividad manteniendo el mismo nivel de esfuerzo.

Patologías

En este proyecto se pretende conseguir un algoritmo de aprendizaje automático que clasifique diversos pacientes según sus patologías. Aquellas en las que nos concentraremos son:

- **Listesis:** es el desplazamiento de una vértebra sobre otra puede ser hacia adelante anterolistesis o hacia atrás retrolistesis (1). Es la condición más rara de las que se tratan en este proyecto, teniendo solo 13 representantes en el conjunto de datos proporcionado; compuesto por un total de 221 pacientes.
- **Estenosis Foraminal y de Canal:** una estenosis hace referencia a un estrechamiento del conducto por donde pasan los nervios ejerciendo presión en estos lo que causa dolor y pérdida de sensibilidad (2). La estenosis Foraminal (estenosis F) es aquella en la que el estrechamiento del espacio entre las vértebras presiona las ramificaciones de la médula espinal (3), mientras que la estenosis de canal (estenosis C) ocurre en el canal por donde pasa la propia médula, presionando esta (4).
- **Artrosis:** Es la más común de las cuatro lesiones en el grupo de estudio y probablemente la más conocida, ya que no solo afecta a la espina dorsal. La artrosis es una patología reumática que desgasta el cartílago articular. Este desgaste produce dolor, al reducir la amortiguación que proporcionan los discos. También causa indirectamente otras patologías, ya que la falta del disco acelera a su vez el desgaste de las vértebras produciendo su deformación y causando estenosis del canal medular. Por último, el desgaste de los discos y consiguiente aproximación de las vértebras hace que los huecos por donde salen las ramificaciones de la médula espinal se estrechen ejerciendo presión en estas, lo que hemos denominado estenosis Foraminal. (5)

También es posible clasificar a los pacientes según sean o no patológicos, indistintamente de cuál sea la patología.

Otras patologías que se tendrán en cuenta, pero cuyo diagnóstico no se incluirá en el programa final son DBE y Ciática. Ello es debido a que ambas son diagnosticadas en consulta y no a partir de la imagen de RMN. A continuación, se explican brevemente

- **DBE:** Corresponde al acrónimo de Dolor Bajo de Espalda, es decir, no es una patología en sí, si no una observación hecha en consulta y un dato para tener en cuenta a la hora de clasificar.
- **Ciática:** La ciática es una condición que supone dolor irradiando del nervio ciático. En primera instancia se incluyó en el grupo de patologías a clasificar ya que una de las causas más comunes es el pinzamiento del nervio por una hernia de disco, estrechamiento de columna o espolón óseo (6). Como el DBE, es también un síntoma, causado por otras de las patologías tratadas en este proyecto, como la Estenosis de Canal. Por tanto, si se diagnostica en consulta no hay necesidad ni razón para clasificar a los pacientes que la sufren según su imagen de la RMN y, es más, podemos usarla para ayudar a la clasificación.

Otras clasificaciones de las que se disponía de información pero que no fueron consideradas para los algoritmos finales son:

- **Discopatías:** no son en sí enfermedades, sino simplemente la pérdida de altura de uno de los discos, en la mayoría de los casos como consecuencia de un desgaste degenerativo. A diferencia de las otras clasificaciones, esta se presenta en la base de datos disponible como el número de discopatías, de una a cinco, que se aprecian en los cinco discos lumbares (7). El hecho de que no sea una clasificación binaria hace que la aplicación del método que se sigue con el resto de las patologías no sea válida y haga falta tratarla en paralelo, si bien, dada la diferencia entre la presentación de estos resultados y los demás y el hecho de que no se apliquen las curvas ROC directamente, como resultado de validación, hace que el estudio de las discopatías haya quedado relegado a un segundo plano y como una posible extensión de futuro.
- **Signos Modic:** Al igual que algunos de los casos anteriores, no es en sí una patología, sino algunos marcadores que se pueden apreciar en las RMN y ayudan al médico a hacer el diagnóstico (8). Estos signos pueden ser útiles para ayudar al diagnóstico del resto de patologías, pero hemos constatado que, en nuestro caso, es difícil detectarlos mediante técnicas de aprendizaje automático y como se miden en la propia RMN no podemos saber de antemano si se encuentran en el paciente a diagnosticar, a diferencia del DBE y la ciática.
- **Cirugía:** indica si el paciente será o no sometido a cirugía. En un principio se tuvo en cuenta como una posible salida de los algoritmos, pero al ver el poco interés de incluirla, ya que es inútil indicar que el paciente puede que se vaya a considerar como objeto de cirugía sin que el médico vea la RMN.

Problemática

Una de las limitaciones de este trabajo se encuentra en la infrarrepresentación de ciertos grupos, ya que en tres de las patologías a tratar: listesis, estenosis C y estenosis F, el porcentaje de representación es de menos de 6, 11 y 13% respectivamente

Esto es un problema porque cuantos menos pacientes afectados vea el algoritmo a la hora de entrenar, será más difícil que aprenda a reconocerlos. Además, esto es doblemente

problemático en este proyecto, porque al tratarse de sanidad, un falso negativo es mucho menos deseable que un falso positivo. Por último, un programa que no detecte bien a los enfermos, pero sí a los sanos podría parecer bueno desde el punto de vista que acierta un elevado número de veces, sin embargo, nos sería inútil. Aquí es donde vemos la importancia de elegir bien el criterio por el que se decide si el algoritmo funciona o no, ya que, por ejemplo, en el caso de la listesis, que padecen solamente 13 de los 221 pacientes, un algoritmo que marcara todos como sanos nos brindaría una exactitud del 94.11% (13), así que, tendremos que fijarnos en otros indicadores que serán la exactitud, precisión y otros derivados de ellos y se explicarán en Material y Método. En última instancia, habrá que elegir cual es el mejor balance para la sensibilidad y la precisión de la clase positiva al ser deseable detectar al mayor número de pacientes enfermos, que luego pueden descartarse manualmente, pero tampoco ser cuestión de marcar a un gran número de sanos erróneamente, pues no se conseguiría la ayuda esperada para los facultativos. Como, en cualquier caso, el balance entre exactitud y sensibilidad es un parámetro importante para el diseño final del algoritmo. se ha implementado este balance de forma fácilmente modificable por quien pretenda utilizar los programas desarrollados.

También hay que tener en cuenta que el conjunto de puntos que tenemos es escaso y no nos proporciona información sobre la densidad ósea, que se puede ver en las RNM y puede ser un factor de importancia a la hora de diagnosticar. Sólo podremos fijarnos en la geometría definida por los puntos mencionados que no el conjunto de las vértebras lumbares, así como algún otro factor como edad, peso, altura, sexo... poder trabajar con este dato podría mejorar mucho los resultados en trabajos futuros.

Objetivo

El objetivo de este proyecto es por tanto desarrollar un algoritmo que pueda clasificar un grupo de pacientes en función de si padecen o no estenosis F, C, artrosis, listesis o cualquier combinación de estas y los que están sanos, para que, en un futuro, junto con otras investigaciones en curso, puedan generarse lecturas automáticas de las RMN y asistir en el diagnóstico al médico, quitándole carga de este trabajo más rutinario para que pueda orientar sus esfuerzos a otras tareas donde haga más falta el factor humano. En el ámbito de este trabajo la clasificación se hará en base a los puntos geométricos correspondientes a las esquinas de las cinco vértebras lumbares, los puntos medios de sus lados y el tejido blando que se puede ver en el plano sagital medio de la RMN. En el siguiente apartado se muestran algunas imágenes para ilustrar a qué puntos nos referimos. En los datos de los que disponemos para este proyecto los puntos han sido marcados manualmente por un profesional, de ahí la escasez de pacientes. Se está investigando para poder determinar estos puntos automáticamente a partir de la RMN en un futuro, e incluso más información como la densidad ósea o puntos en más planos que el sagital medio.

Material y Métodos

Datos Disponibles

Los datos de los que se dispone para desarrollar los modelos están recogidos en una tabla donde cada fila corresponde a un paciente y en cuanto a las columnas la primera es el índice, seguido de ocho “patologías” distintas, una casilla que determina si el paciente será sometido o no a cirugía, una que es combinación lineal de las seis primeras patologías, una que marca si la RMN es o no patológica, varias con información sobre el paciente como sexo, edad, peso y por último 132 casillas con números que representan las coordenadas X e Y de los siguientes puntos en la RMN, en esta página se muestra el plano sagital medio de una RMN con los puntos marcados junto con dos esquemas que contienen solo los puntos que se han utilizado en la confección de este proyecto, uno sano y otro que padece todas las patologías:

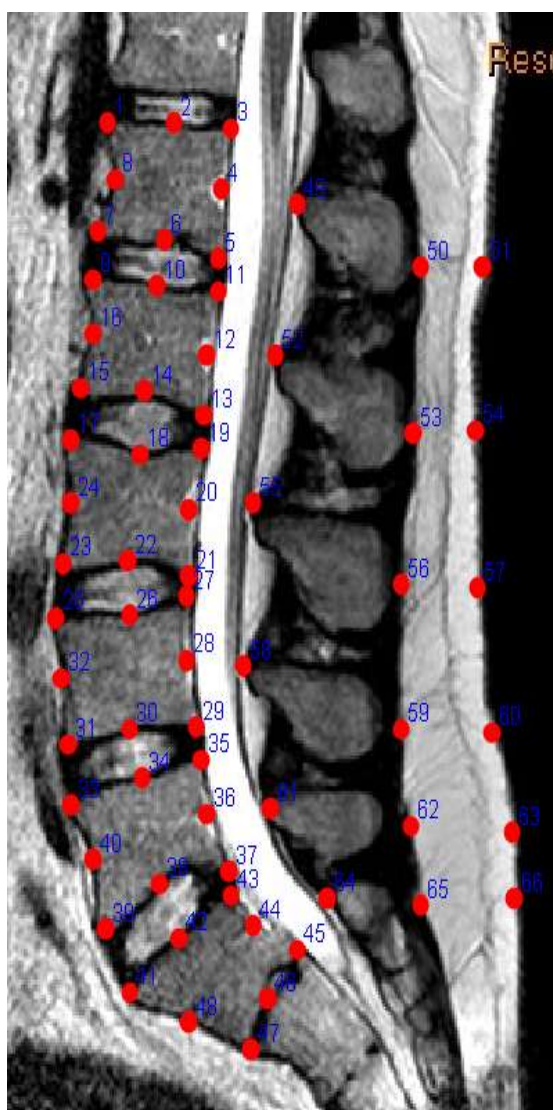


Ilustración 3 RMN con los marcadores

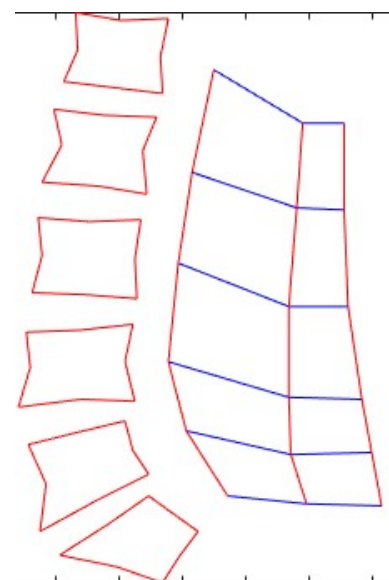
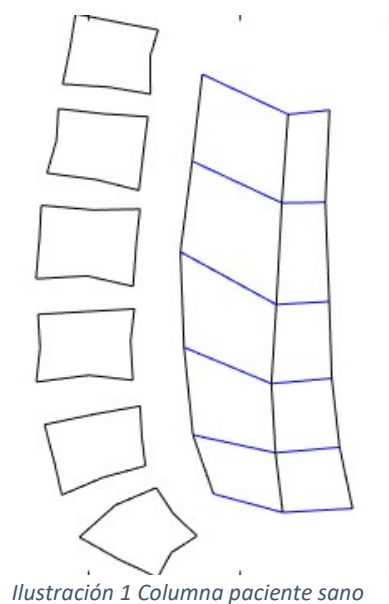


Ilustración 2 Columna paciente con todas las patologías

159	0	0	0	0	0	0	0	0	0	0	NORMAL	HOMBRE	38	179	69	21,53	88	0,49	NORMAL	NO	159	15,89376	24,039312	17,54936	23,906864	19,072	
160	5	1	0	1	0	0	1	0	1	0	7	PATOLOGICA	MUJER	70	147,5	63,2	29,05			SOBREPESO	SI	160	15,865318	22,864723	17,265199	22,66474	18,66
161	2	1	0	1	0	0	1	0	1	1	5	PATOLOGICA	HOMBRE	40	186,1	92,3	26,83			SOBREPESO	SI	161	19,129411	24,568304	20,995695	24,528304	22,728
162	3	1	1	1	1	0	1	1	1	1	7	PATOLOGICA	HOMBRE	74	159,5	73,8	29,01			SOBREPESO	SI	162	16,531928	25,99779	18,465097	25,531163	20,98
163	2	1	0	0	0	0	1	1	1	1	3	PATOLOGICA	MUJER	62	163,7	60,8	22,69			NORMAL	NO	163	21,918493	23,044212	24,441526	22,731117	21,03
164	2	1	0	0	0	0	1	1	1	1	3	PATOLOGICA	MUJER	57	152,1	69,7	30,13			OBESIDAD	SI	164	21,132805	20,79948	22,26611	20,466155	23,399
165	1	0	0	0	0	0	0	1	0	1	0	PATOLOGICA	HOMBRE	53	179	104	32,46			OBESIDAD	SI	165	15,278514	25,09245	16,91417	24,572014	18,438

Técnicas de clasificación

Algoritmos de Aprendizaje Automático

- **Support-Vectors Machine:** es un algoritmo que distribuye los elementos en un espacio n -dimensional, siendo n la longitud del vector asociado a cada punto, y posteriormente busca un hiperplano que separe los que pertenezcan a una clasificación de los de la otra con el mayor margen posible, de ahí surge su nombre ya que estos vectores son los más cercanos al hiperplano y a partir de los cuales este se define. (9)
- **Discriminant Analysis:** Con los datos de entrenamiento, *Discriminant Analysis* calcula la media y varianza, asumiendo una distribución normal de los mismos en un espacio multidimensional. Al recibir nuevos datos calcula la probabilidad de pertenencia a cada grupo asignándole aquel que tenga una probabilidad mayor (10).
- **Decision Trees:** Devuelve un árbol de decisión binaria donde los nodos son valores de los puntos de la tabla de predicción (11). No da buenos resultados por sí solo, sin embargo, se tuvo en cuenta para combinarlo con los demás y así tener tres programas diagnosticar en cada caso siguiendo a la mayoría, ya que al ser las clases binarias dos siempre darán la misma solución.

Para evaluar si las predicciones en base al modelo eran correctas o no, se utilizó una herramienta llamada matriz de confusión que recoge cómo se han clasificado los datos frente a su clase (enfermo o sano) real para cada una de las predicciones tal que:

PREDICHO REAL	SANOS	ENFERMOS
SANOS	SANO VERDADERO (A)	FALSO POSITIVO (B)
ENFERMOS	FALSO NEGATIVO (C)	ENFERMO VERDADERO (D)

7

Inicialmente se evaluó el funcionamiento del programa únicamente con el porcentaje de aciertos, ya que, en una primera instancia, parece razonable y es intuitivo. Sin embargo, esta no es la única forma de determinar si la clasificación es buena y en algunos casos tampoco la mejor, como se ha discutido ya en el apartado de Problemática.

Es por eso que se han utilizado otros indicadores como la sensibilidad y la precisión. La sensibilidad indica qué fracción de los elementos de una clase ha sido clasificada como tal, mientras que la precisión indica cuantos de los elementos identificados como una clase realmente pertenecen a ésta. En problemas donde la no-detección de una clase en particular tiene un alto coste, la sensibilidad de la clase en cuestión debe ser el indicador primordial. No obstante, no puede ser el único indicador que se considere, ya que un programa que clasificara a todos los datos en ese grupo particular sería considerado ideal según dicho criterio.

Con la nomenclatura indicada en la matriz de confusión, la exactitud del algoritmo sería determinada como:

$$\frac{A + D}{A + B + C + D}$$

Ecuación 1 Exactitud

La sensibilidad de la clase positiva o TPR, en este caso serían los enfermos:

$$\frac{D}{C + D}$$

Ecuación 2 TPR

La precisión de la clase positiva:

$$\frac{D}{B + D}$$

Ecuación 3 Precisión

Para la clase negativa (sanos) bastaría reemplazar D por A, B por C y viceversa (13).

La curva ROC (14) (*Receiver Operating Characteristics*) es una representación gráfica de la sensibilidad frente a la especificidad que nos permite examinar cuanta precisión sacrificamos a costa de conseguir más sensibilidad.

En el eje de abscisas se representa la sensibilidad, o TPR (*True Positive Ratio*), ya definida, que será igual a cero cuando no se detecte ningún enfermo y uno cuando todos los enfermos sean detectados.

En el eje de ordenadas veremos la complementaria a la sensibilidad de la otra clase o FPR (*False Positive Ratio*):

$$1 - \frac{A}{B + A} = \frac{B}{B + A}$$

Ecuación 4 FPR

El FPR será cero cuando no haya ningún sano clasificado como enfermo y uno cuando todos los sanos sean clasificados como enfermos.

Una vez asignada una probabilidad de pertenencia a la clase en cuestión, cada punto de la curva ROC representa un umbral a partir del cual consideramos que cada paciente pertenece a una u otra clase.

En la ilustración 2 se muestran varios ejemplos de curvas ROC. La forma ideal de esta curva, que tendría un algoritmo que separa los puntos a la perfección sería la verde, mientras que la peor forma que puede tener está representada por la curva roja, que significaría una separación nula de las clases, la curva azul separaría las clases perfectamente, pero estas estarían invertidas (14).

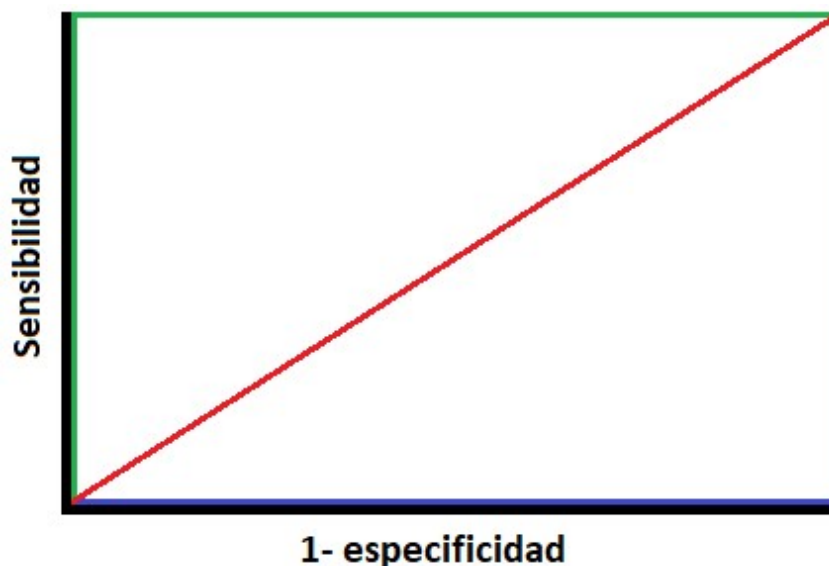


Ilustración 6 Ejemplo Curvas ROC

Comparación de Algoritmos

El procedimiento con el que se ha comparado y juzgado qué algoritmos eran mejores ha consistido en ejecutar un programa donde se entrenaba y predecía con distintos algoritmos un gran número de veces para conseguir que los distintos indicadores convergiesen. En las primeras fases del proyecto el principal indicador fue la eficacia, simplemente dividiendo los aciertos entre el número de pruebas, usando la función *confusionmat* (g1, g2) de Matlab.

Más adelante se cambió el criterio a la sensibilidad de las clases, precisión y especificidad.

La aproximación que se tomó para una primera evaluación y familiarizarse con las herramientas disponibles fue un sencillo programa que devolvía la exactitud de los algoritmos utilizados: *Decision Trees* (en adelante Trees), *Discriminant Analysis* (en adelante DA) y *Support-Vectors Machine* (en adelante SVM), usando en cada caso diferentes porcentajes de entrenamiento, entre el 20 hasta el 95 para ver cómo mejoraba el programa y si llegaba a un máximo de eficacia o si no alcanzaba este máximo estimar cuánto podía mejorar con el uso de más datos.

Purga de lista de entrenamiento

A raíz de intentar darle explicación a unos resultados extraños usando DA para porcentajes de entrenamientos comprendidos entre 45 y 65, pensó que podía haber pacientes que hacían que la eficacia de los programas bajase, por ejemplo, por no ser buenos representantes de su clasificación, ya sea que tienen una Listesis muy ligera o no son artríticos pero la distribución de

los puntos puede dar a entender que sí, aunque la densidad ósea medida en la RMN muestre que no lo son.

Por tanto, la primera mejora fue desarrollar un programa que detectase estos supuestos pacientes poco representativos. La forma de implementarla fue la siguiente: Primero se guardaban los conjuntos de entrenamiento en grupos de cien, se guardaban el peor y el mejor conjunto de entrenamiento de dichos cien y posteriormente se contaba cuántas veces aparecía cada paciente en cada grupo. Se observaron unas claras tendencias, donde unos pacientes se repetían mucho más que otros en una tabla cualquiera, siendo estos mismos los que menos aparecían en la tabla complementaria de una misma enfermedad. Es decir, los que más aparecían, por ejemplo, en el grupo de estenosis F con malos resultados eran de los que menos aparecían en el grupo bueno de estenosis F. Además, no solía haber correlación entre patología y patología. Otra tendencia que se observó, pero no se supo interpretar entonces, fue que estos pacientes solían pertenecer al grupo positivo de las respectivas patologías. Un análisis más detallado de estos aspectos se incluye en el apartado de Resultados.

Algoritmos en Tándem

La primera mejora implementada fue el uso de dos algoritmos conjuntamente, en el cual ambos generaban un modelo independientemente. De esta manera cuando ambos algoritmos indicaban que un paciente estaba sano, era más probable que lo fuese y viceversa y en el caso en que no estuviesen de acuerdo se determinaba al paciente “confuso”, para posteriormente decidir un criterio que determinase la clase de estos. La implementación de esta idea hizo que retomara el uso de Trees para poder hacer usar una técnica, en la cual los tres algoritmos diagnostican y el paciente es clasificado en la clase que determinan al menos dos de los algoritmos, esto se llama *minority report*. Dado que las clases son binarias siempre habrá por lo menos dos que estarán de acuerdo.

Scores SVM

Para decidir la clasificación de los pacientes “confusos” inicialmente se planteó utilizar el *score* proporcionado por SVM, que es la distancia de cada punto al hiperplano que separa las clases (9), dándole la razón a este si el valor absoluto del *score* era mayor que un umbral y a DA en el caso contrario, creando una tabla con la eficacia para distintos valores de dicho umbral.

Generación Tabla Probabilidad

Una de las últimas etapas hacia la elaboración del programa final fue la generación de una tabla que, en teoría, si los resultados eran buenos, debía parecerse a la tabla de pacientes-patologías (zona coloreada de la ilustración 6). Para ello se debía contar el número de veces que cada paciente era diagnosticado ordenándose entonces de forma que el resultado fuese una matriz que indicaba cuán probable era que cierto paciente fuese diagnosticado con cierta patología. La manera de hacer esto fue con tres tablas, una de puntuación, una de conteo, y la propia tabla de probabilidad. Después de determinar el grupo de predicción, se sumaba uno a cada casilla de la tabla de conteo correspondiente a los pacientes que se encuentran en el grupo de predicción, y después de hacer el diagnostico se añadía uno en la tabla de puntuación si el paciente era diagnosticado con enfermedad, cero si era clasificado sano, y en los casos donde el diagnostico era confuso (los algoritmos no se ponían de acuerdo) un valor entre cero y uno que era dependiente del método en cuestión, por ejemplo, si usamos SVM con DA y aplicamos un criterio más “pesimista” para los casos confusos y así detectar más casos a costa de tener algún falso positivo más, se añadiría 0.75, mientras que si teniendo en cuenta que los umbrales que utilizaremos finalmente serán más bajos de 0.5, podemos añadir 0.25 a la casilla

correspondiente a ese paciente-patología. Tras repetir el programa un elevado número de veces, dividir la tabla de puntuación entre la de conteo daría la probabilidad de que cada paciente padezca cada patología. En esta fase se calculaban todas las patologías a la vez, con sólo un método por ejecución. El tiempo de ejecución del programa era elevado y por tanto limitaba el número de métodos que se podían probar, sin embargo, me brindó unos resultados muy interesantes. Al final del proyecto se consiguió atenuar el problema de la demora en los cálculos eliminando clasificaciones, entre otras cosas.

Curvas ROC

En una fase ya próxima al final del trabajo se cambió el método: se analizaría para cada patología diferentes métodos, en vez de analizar un mismo método para todas las patologías de vez. Así podía ajustar mejor el método para cada patología, y era posteriormente más cómoda la comparación de los métodos. Asimismo, se dejó de utilizar Excel para el procesamiento posterior de los resultados porque las ventajas, eran sobrepasadas por los inconvenientes y limitaciones de una hoja de cálculo para esta tarea. En lugar de Excel se utilizó Matlab.

En este caso utilicé dos programas, el primero de ellos, similar al del apartado anterior, generaba una matriz, en este caso de dimensiones (pacientes, métodos, %entrenamiento) para cada una de las cinco clasificaciones finales, descritas en el apartado “introducción-patologías”. Una vez realizada dicha matriz se recorre comparándola con un umbral que variaba entre cero y uno en intervalos de 0.0005, por tanto, 2001 intervalos produciendo dos matrices de dimensiones (2001, métodos, %entrenamiento) dimensiones, en las que, una de ellas guardaba la evolución del TPR (sensibilidad, que fracción de los enfermos se han detectado) a medida que aumentaba el umbral y la otra guardaba la evolución del FPR (que fracción de los sanos se han clasificado como enfermos) y representaba estas curvas para el binomio método - %entrenamiento deseado.

Finalmente, para elegir la “mejor” curva, en vez de utilizar el criterio AUROC (área bajo la curva ROC, ilustración 2), como medida de la calidad del programa separando las clases, se buscó el punto; ya que lo que nos interesa para el proyecto es el umbral óptimo en el mejor de los métodos posibles, que me devolvía una mayor puntuación. Siendo esta puntuación, una que elegí de forma más o menos arbitraria pero que con la opinión de un profesional de la salud podríamos cambiar con poco esfuerzo, como se ha mencionado con anterioridad este criterio está implementado de forma que es fácilmente modificable:

$$puntuación = y - 0.5x$$

Por tanto, cada punto me dará un valor distinto y el que me dé el mayor valor sería el elegido.

La razón de este indicador y no otro es que, por un lado, mayor “y” mejor, porque significa una mayor detección, y menor “x” mayor precisión (menos FPR). Como ya se ha explicado, por la naturaleza del problema es más importante la sensibilidad que la precisión.

Para calcular cada puntuación, y la máxima dentro de todas ellas, calculé una nueva matriz donde punto a punto, de las curvas, se calculaba la puntuación correspondiente, después el máximo en cada curva, guardando cual era el umbral correspondiente y el máximo entre los máximos. Finalmente, el programa me devuelve un vector con la puntuación, sensibilidad, 1-especificidad, umbral, método y %entrenamiento.

También con esta metodología se estudió añadir nuevos datos al vector a partir del cual se genera el modelo y predice, como la altura, peso, sexo, edad y existencia o no de ciática y DBE

en el paciente, en vez de guiarnos solamente por las coordenadas de los puntos marcados. Lo cual mejoró los resultados.

Este método fue desarrollado independientemente, pero después de finalizar el trabajo busqué si ya existía algo parecido y encontré que así era (15). Sin embargo, la forma de aplicarlo, entrenando con distintos subgrupos de un grupo limitado de entrenamiento para conseguir probabilidades distintas sí que parece ser innovador.

Separación grupo de verificación

Este apartado es la fase final del proyecto y la prueba de que el método es eficiente a la hora de detectar pacientes con lesiones a partir de los puntos extraídos del plano sagital medio de una RMN usando algoritmos de aprendizaje automático.

Se ha escogido un grupo que será con el cual verificaremos que el algoritmo elegido funciona, se ha intentado que la proporción de pacientes en cada enfermedad sea representativa del total en el grupo de verificación, así pues tenemos un grupo de 40 pacientes: 3 listésicos (7.5%), 12 artrósicos(30%), 7 con estenosis F (17%) y 5 con C(12%) y otros con discopatías o patologías no incluidas en el estudio para un total de 29 patológicos (72.5%) mientras que en el grupo total había unos porcentajes respectivos de 5.9, 24, 12, 11 y 65%.

Una vez este grupo ha sido separado se procede como en el punto anterior, eligiendo un conjunto de métodos y un rango de porcentajes en los que probarlos. Se repite un número que se considere adecuado de veces para que los resultados converjan y una vez tenemos la matriz de probabilidades calculada por el programa, de dimensiones (pacientes de modelado, métodos, porcentajes), la introducimos en el programa que dibujaba las curvas ROC y devolvía el punto óptimo de entre los disponibles, informando también de qué umbral se debe utilizar para obtener esa puntuación. Por último, se ejecuta un programa cuyo objetivo es diagnosticar al grupo de verificación repetidas veces con distintas variaciones del grupo de entrenamiento para obtener de nuevo probabilidades estables. Para terminar, se transforman estos umbrales en clasificaciones sano/enfermo aplicando el umbral hallado en el programa anterior. Para juzgar los resultados se utiliza la matriz de confusión y se comprueba que sean parecidos a los obtenidos en el grupo de creación del modelo.

Los vectores discriminantes que se han utilizado han sido la geometría, existencia de DBE y ciática, peso y edad, excepto para el caso de patología donde no se han incluido DBE y Ciática.

Resultados

Primeras Tablas

Las tablas 1, 2 y 3 recogen los datos de la eficacia de cada algoritmo, SVM, DA, y Trees respectivamente. Cada fila representa una clasificación (patología), mientras que cada columna indica el porcentaje de pacientes que se usaron como entrenamiento, el código de colores pretende ayudar a la lectura de estas siendo un degradado de rojo oscuro (menor valor) a verde oscuro (mayor valor) pasando por el blanco

	25	35	45	55	65	75	85
MODIC	81,01	81,7	82,03	81,44	81,26	81,36	81,06
LISTESIS	90,34	91,4	91,25	91,72	91,56	91,47	90,94
ARTROSIS	74,74	75,84	77,05	77,22	77,57	76,91	77,55
ESTEN_F	84,82	85,74	85,78	86,67	86,94	86,93	86,27
ESTEN_C	86,25	86,58	86,93	87,7	87,08	87,67	88,88
DBE	66,7	69,04	70,38	71,46	71,53	71,87	71,94
CIATICA	68,3	69,55	70,11	70	71,04	70,05	70,73
CIRUGIA	69,23	70,15	71,16	70,8	71,52	72,18	73,45
PATOLOGIA	80,83	85,92	86,98	89,43	90,19	92,02	91,76

Tabla 1 Eficacia SVM

	15	25	35	45	55	65	75	85	95
MODIC	80,47	78,26	75,28	70,7	62,96	62,16	70,89	73,88	77,55
LISTESIS	91,78	91,14	90,03	85,69	73,21	73,78	82,38	86,15	87,82
ARTROSIS	71,57	72,81	70,8	67,34	60,77	60,83	66,45	71,15	72,91
ESTEN_F	83,55	83,07	80,01	75,54	65,69	65,18	71,84	74,61	79,09
ESTEN_C	85,76	83,87	82,22	78,15	66,87	65,99	73,84	77,97	78
DBE	61,44	64,9	65,45	64,6	58,62	58,84	66,65	70,58	73,91
CIATICA	64,21	65,48	63,46	62,12	56,79	56,01	61,2	62,36	64,36
CIRUGIA	65,26	66,43	64	60,65	56,47	57,13	62,35	64,82	65,82
PATOLOGIA	74,31	79,93	81,36	78,23	68,45	69,26	79,53	82,97	85,91

Tabla 2 Eficacia DA

	15	25	35	45	55	65	75	85	95
MODIC	77	75,3	75,7	74,3	75	75,2	75,1	74	70,5
LISTESIS	90,1	90,3	90,4	90,4	90,8	90,8	90,5	91,4	89,5
ARTROSIS	65	66,6	67,2	67,7	67,8	68,3	68,5	69	68,8
ESTEN_F	81	80,2	80,1	79,6	79,6	80,8	80,9	80,8	80,2
ESTEN_C	82,3	82,6	81,8	80,2	81,1	81,5	81,3	81,9	82,4
DBE	53,1	53,1	53,5	54,4	54,5	54,2	53,8	54,2	56,5
CIATICA	50,8	51,7	51,5	51,6	51,4	52	52,9	52,2	53,3
CIRUGIA	59,4	59,4	58,5	59,7	58,9	59,9	58,5	58,4	58,2
PATOLOGIA	59,6	61,4	61,5	61,6	62,1	63	62,8	63,1	61,1

Tabla 3 Eficacia Trees

A partir de ellas se puede observar la eficacia y cómo varía esta, según se utilicen distintos porcentajes de entrenamiento. La conclusión obtenida fue que SVM daba los mejores resultados, si bien DA da menos falsos negativos que SVM, cosa que no se aprecia en las tablas, pero se notó al ejecutar el programa, Trees parecía peor en todos los aspectos. Fue la diferencia entre falsos positivos de SVM y DA lo que dio lugar a la idea de usarlos juntos. Dados estos resultados se descartó Trees por el momento. También es interesante anotar que DA y Trees eran parecidos en cuanto a velocidad mientras que SVM sólo en algunas patologías y a bajos porcentajes de entrenamiento alcanzaba velocidades similares, siendo mucho más lento en porcentajes mayores.

Se puede observar un fenómeno extraño que es la caída en la eficacia de DA. Al tratar de explicar esta tendencia se consideró que estuviese relacionada con los pacientes usados para el entrenamiento. De ahí surgió la idea de filtrar los pacientes que daban “peores” modelos de la lista de entrenamiento.

Por último, se observa una mejoría, pequeña en general, para SVM, al usar más pacientes en el grupo de entrenamiento sin embargo esto conllevaba inconvenientes como el aumento del tiempo para cada predicción durante la fase de aprendizaje. En consecuencia, se decidió que esta información no era suficiente para fijar un porcentaje de entrenamiento, por lo que se continuó el estudio conservando la variación de dicho parámetro.

Mejoras

En este apartado se presentan algunos resultados tanto del método de filtrado de pacientes como de SVM y DA funcionando en conjunto

Las tablas que se ven a continuación corresponden al número de veces que cada paciente ha aparecido en los grupos de entrenamiento “buenos” y “malos”. Cada casilla corresponde a un paciente, ordenados de arriba abajo y de izquierda a derecha, siendo la esquina superior izquierda el paciente 1, la inferior izquierda el paciente 17 y la inferior derecha el 221. Para la confección de cada tabla se han usado 200 grupos, y hay cuatro tablas por patología, dos de grupos que dieron buenos resultados, por ejemplo, la tabla 4, y dos grupos cuyos resultados fueron peores, como en la tabla 5. Cada grupo consta de 33 pacientes.

40	33	25	30	23	42	28	25	87	25	30	22	26
28	22	28	74	36	31	28	22	24	24	34	21	33
29	31	22	26	37	3	29	37	24	19	26	38	25
28	19	36	20	32	36	21	41	35	34	34	19	28
33	24	26	24	30	28	19	32	34	26	25	25	30
32	30	24	27	25	26	30	30	33	35	30	26	25
28	34	21	21	33	35	25	25	19	36	27	26	33
35	32	32	32	31	52	25	30	32	31	29	37	24
29	30	28	30	24	25	22	18	36	52	39	32	31
28	32	31	27	23	28	25	20	32	36	25	23	38
31	27	38	34	41	27	31	35	32	24	42	66	24
21	23	31	31	35	34	17	25	32	32	27	35	33
22	32	48	31	26	41	31	37	20	22	43	48	30
32	21	39	21	26	33	31	27	27	27	24	21	18
25	25	28	40	20	39	16	19	27	43	38	52	39
34	30	27	33	40	29	28	28	28	27	22	24	34
25	22	28	33	35	29	8	36	33	25	26	31	20

Tabla 4 Mejores Listesis

30	27	18	27	35	33	15	31	27	31	31	26	23
18	24	34	44	18	22	24	25	26	29	25	30	21
31	38	35	39	16	87	30	22	20	28	22	29	26
34	25	61	35	24	29	36	30	55	22	36	40	30
28	23	38	35	25	35	28	24	27	30	28	29	20
26	28	32	37	35	40	23	22	26	36	26	21	28
29	29	28	28	23	25	24	36	39	34	37	35	29
32	29	36	22	27	21	28	38	29	20	30	22	34
34	30	30	34	21	28	32	38	28	47	27	24	33
23	23	33	26	19	28	31	28	26	25	26	28	21
26	29	29	22	28	30	22	25	23	22	27	38	31
67	25	21	34	30	26	35	33	24	28	39	27	22
27	27	41	35	33	27	24	25	22	27	25	30	31
37	24	29	27	32	32	30	27	35	23	30	29	29
19	25	17	29	25	25	30	30	37	25	28	73	21
39	31	28	27	27	30	24	24	32	18	32	22	30
38	33	28	21	26	27	118	33	39	32	33	34	38

Tabla 5 Peores Listesis

El filtrado de los pacientes malos para el entrenamiento, cada vez que aparecían aleatoriamente en el grupo de entrenamiento, mejora en la eficacia de los programas, como se observa en la tabla 6 para el SVM, en comparación con la tabla 1

	15	25	35	45	55	65	75	85	95
M	83	83,8	84,3	84,3	85	85,4	84,8	85,1	85,4
L	93,1	93,5	93,3	93,8	93,5	93,8	93,7	94,3	93,7
A	77,2	79,1	80	80,9	81,7	82	82,7	82,9	84,1
F	86,9	87,2	88,1	88,7	88,8	89,2	89,5	89,6	89,2
C	87,6	87,4	87,7	87,8	87,9	87,9	88,3	88,1	87,5
D	65,2	67,7	70,7	72,9	73,8	75	76,6	77,2	81,5
Z	67	70,9	71,3	72,9	73,7	74	74,7	74,2	76,2
G	69,8	71,1	71,7	72,9	73,5	73,7	74,5	74,2	73,8
P	76,2	82,4	85,6	87,8	89,9	91,8	92,6	93,3	92,8

Tabla 6 SVM con filtro

Mientras que para algunas de las clasificaciones la variación es nula o está dentro de lo que podría ser normal debido a la aleatoriedad de los datos, en otras como la Artrosis pasa de 77,5 a 84 o DBE de 72 a 81,5. En el caso de la Listesis llegamos a porcentajes tan altos como 93,7.

Sin embargo, una vez introducidos los indicadores de sensibilidad y precisión se vio que purgando estos pacientes la sensibilidad de la clase positiva bajaba, diagnosticando menos como enfermos y por tanto diagnosticando menos sanos como enfermos, lo cual, debido a la falta de balance en los datos hacía que mejorase la eficacia, aquí se ve por tanto la importancia de fijarse en el indicador de la sensibilidad.

En cuanto a la mejora obtenida al hacer cooperar a los algoritmos (Tabla 7), se realizaron pruebas con un 75% de pacientes en el grupo de entrenamiento y se observó, de izquierda a derecha, la eficacia en aquellos pacientes donde los algoritmos coincidían (éxito), el porcentaje de pacientes que no eran confusos (acuerdo), la precisión de la clase sana (seguridad), y el porcentaje de sanos en el grupo de confusos (sanos). También se muestran las desviaciones típicas de los tres primeros indicadores.

	éxito	Acuerdo	Seguridad	Sanos			
modic	85,13	75,64	91,09	79,30	3,48	3,55	1,92
listesis	94,05	84,89	95,61	90,92	3,10	3,64	1,45
artrosis	84,03	71,57	89,24	72,41	3,62	3,64	1,94
estef	89,06	76,69	92,44	82,76	3,21	3,36	1,75
estec	90,56	78,37	93,27	83,09	3,10	3,19	1,71
dbe	78,81	71,69	89,78	39,29	3,57	3,59	1,98
ciatica	72,37	69,55	85,41	53,17	3,69	3,77	2,25
cirugia	76,22	69,65	86,37	69,23	3,51	3,60	2,17
Pat	95,15	78,26	97,83	30,19	3,58	3,46	1,06

Tabla 7 Resultados SVM+DA 75% entrenamiento

Aquí se aprecia que la eficacia es elevada aún sin haber filtrado. Fue a partir de estos datos cuando se decidió intentar recrear la tabla de casos, para ver si los fallos se daban en algunos pacientes en particular o indiscriminadamente y si pasaba lo mismo con los confusos.

Antes de continuar con esa perspectiva se estudió la posibilidad de utilizar los scores de SVM para determinar la clasificación de los pacientes confusos, este método no dio resultados concluyentes. En la ilustración 6, el eje de abscisas presenta la exactitud mientras que el de ordenadas recoge el score mínimo (en valor absoluto) a partir del cual se califican los confusos según SVM

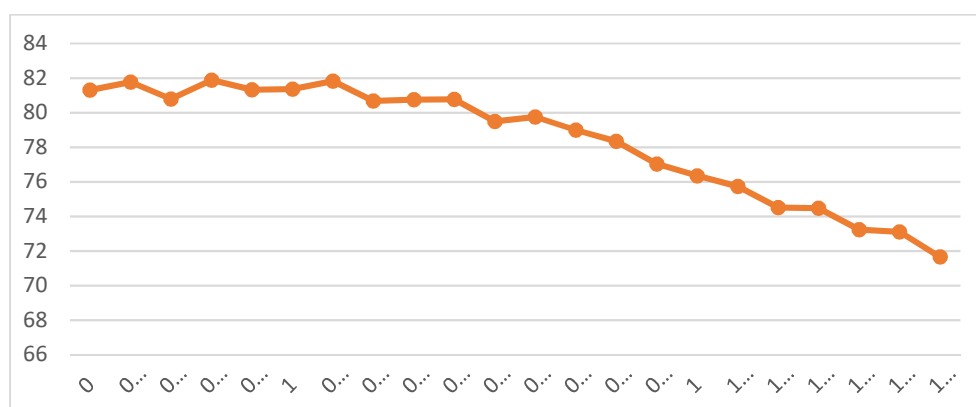


Ilustración 7 Variación de la eficacia respecto al score de SVM

Podemos ver que la variación es prácticamente de derivada negativa a lo largo de toda la gráfica. Lo que sería de esperar si el score no fuese representativo sería una línea recta entre las dos eficacias, mientras que una curva que apoyara la hipótesis de que el uso del *score* es un buen indicador ascendería hasta un máximo destacado. Este caso se asemeja más a lo primero.

Tabla probabilidades

Las tablas completas, están organizadas en páginas de Excel, contienen la tabla de probabilidades, generada por el programa, una tabla de 0 y 1 función de la primera tabla y los umbrales que se elijan, la tabla real y finalmente una tabla que surge de la comparación de estas dos últimas. Para disponer de los resultados de una forma más compacta y fácil de analizar se contaron los pacientes correctamente clasificados y los incorrectos, también marcando para cuales el algoritmo estaba “seguro”, en función de cómo de cerca estaban las probabilidades de cero y uno.

En la ilustración 7 se muestra una tabla que recoge los conteos mencionados. Los números subrayados en rojo corresponden a errores y los verdes a aciertos, el signo menos indica que el diagnóstico ha sido negativo, los valores 1 y 0,1 indican que la probabilidad estaba cerca de 0 o 1 mientras que 0,25 y 0,75 indica que la probabilidad tenía un valor intermedio. Las letras corresponden una a cada enfermedad siendo F y C estenosis F y C, Z ciática y G cirugía, el resto responden a sus iniciales, P2 es 1 si cualquiera de las otras enfermedades es 1 y 0 en caso contrario y sus resultados son contrastados con aquellos de la columna de patológicos en la tabla real. Por último, se pueden ver el total de falsos negativos, correctos y falsos positivos a la derecha. Correctos positivos es la suma de aquellos que han sido correctamente diagnosticados como enfermos mientras que verdaderos positivos es el total de pertenecientes a cada clase.

		M	L	A	F	C	D	Z	G	P	P2		
	-1	7	10	11	7	7	0	0	4	0	13	46	
	-0,75	21	3	20	18	15	21	29	34	7	0	168	
	-0,25	40	2	59	23	27	48	57	74	39	0	369	TFN 214
	-0,1	144	206	109	170	168	22	41	67	24	70	951	Tcorr 1688
0,9	0,1	0	0	5	0	0	24	11	2	82	131	124	TFP 87
0,1	0,25	5	0	16	2	2	80	60	24	55	0	244	
	0,75	4	0	1	1	2	24	22	16	13	0	83	
	1	0	0	0	0	0	2	1	0	1	7	4	
Correctos positivos		5	0	21	2	2	104	71	26	137	131	368	
Verdaderos positivos		33	13	52	27	24	125	100	64	144	144	582	
deteccion enfermos		0,152	0	0,404	0,074	0,083	0,832	0,71	0,406	0,951	0,91	0,632	
Correctos negativos		184	208	168	193	195	70	98	141	63	70	1320	
Verdaderos negativos		188	208	169	194	197	96	121	157	77	77	1407	
deteccion sanos		0,979	1	0,994	0,995	0,99	0,729	0,81	0,898	0,818	0,909	0,938	

Ilustración 8 Resultados de la Tabla de Probabilidades con umbrales 0.5

En la tabla 8, simplemente se habían redondeado las probabilidades. Tras aplicar umbrales distintos, resaltados en azul (Tabla 9), para cada enfermedad se ven los siguientes resultados:

		M	L	A	F	C	D	Z	G	P	P2		
	-1	4	3	11	4	3	0	0	4	0	5	29	
	-0,75	0	0	0	0	0	21	18	16	11	0	66	
	-0,25	0	0	13	0	0	48	47	51	47	0	206	TFN 95
	-0,1	139	182	109	154	156	22	41	67	24	58	894	Tcorr 1587
0,9	0,1	3	7	5	3	4	24	11	2	82	139	141	TFP 307
0,1	0,25	26	3	36	20	17	80	71	42	51	0	346	
	0,75	44	2	47	24	29	24	32	39	5	0	246	
	1	5	24	0	16	12	2	1	0	1	19	61	
		0,08	0,01	0,13	0,05	0,05	0,5	0,4	0,3	0,6			
	Correctos positivos	29	10	41	23	21	104	82	44	133	139	487	
	Verdaderos positivos	33	13	52	27	24	125	100	64	144	144	582	
	deteccion enfermos	0,879	0,769	0,788	0,852	0,875	0,832	0,82	0,688	0,924	0,965	0,837	
	Correctos negativos	139	182	122	154	156	70	88	118	71	58	1100	
	Verdaderos negativos	188	208	169	194	197	96	121	157	77	77	1407	
	deteccion sanos	0,739	0,875	0,722	0,794	0,792	0,729	0,727	0,752	0,922	0,753	0,782	

Ilustración 9 Resultados de la Tabla de Probabilidades con umbrales variables

Si nos guiásemos estrictamente por la eficacia, a la derecha, podríamos decir que la aplicación de dichos umbrales empeora el programa, pero viendo el aumento de la detección de enfermos y la relativa pequeña perdida en la detección de sanos se puede afirmar que es una mejora.

Repetición de las tablas iniciales

Fue, por tanto, al observar los datos que se han presentado en el punto anterior, donde se replanteo el proyecto guiándose por la detección, o sensibilidad y la precisión, o más adelante el FPR (1-especificidad). En las tablas 8, 9 y 10, SVM, DA y Trees respectivamente, una para cada patología, donde cada columna responde a un porcentaje de entrenamiento y cada fila a un marcador, siendo estos: *precisión* (p0 0 hace referencia a la clase sana y 1 a la enferma, p1), *recall* (r0, r1), y *f1* (precisión sensibilidad y f1 es la media armónica entre la precisión y la sensibilidad) de la clase sana y enferma.

		20	35	50	65	80	95			20	35	50	65	80	95			20	35	50	65	80	95
M	p0	0,8817	0,8857	0,8873	0,892	0,8894	0,8889	F	p0	0,8998	0,9105	0,9107	0,9146	0,9226	0,9253	Z	p0	0,7063	0,7311	0,7512	0,7469	0,7601	0,7718
M	r0	0,9186	0,9226	0,9258	0,9286	0,9298	0,9358	F	r0	0,9498	0,9526	0,957	0,9611	0,9585	0,9524	Z	r0	0,7392	0,7592	0,7581	0,7692	0,7724	0,773
M	f0	0,8998	0,9037	0,9061	0,9099	0,9092	0,9117	F	f0	0,9241	0,9311	0,9333	0,9373	0,9402	0,9387	Z	f0	0,7224	0,7449	0,7547	0,7579	0,7662	0,7724
M	p1	0,3938	0,4229	0,4338	0,4567	0,4445	0,4842	F	p1	0,397	0,4845	0,5203	0,5704	0,5723	0,5605	Z	p1	0,6658	0,6948	0,6987	0,7103	0,7151	0,7189
M	r1	0,3001	0,3226	0,3259	0,3479	0,3271	0,34	F	r1	0,2383	0,3223	0,3318	0,3651	0,4082	0,4411	Z	r1	0,6284	0,6626	0,6908	0,6846	0,7009	0,7175
M	f1	0,3406	0,366	0,3722	0,3949	0,3769	0,3995	F	f1	0,2978	0,3871	0,4052	0,4452	0,4765	0,4937	Z	f1	0,6466	0,6783	0,6947	0,6972	0,708	0,7182
L	p0	0,9454	0,949	0,9483	0,9525	0,9492	0,9478	C	p0	0,9087	0,9117	0,9137	0,9096	0,9099	0,9135	G	p0	0,7722	0,7821	0,788	0,7909	0,8008	0,8002
L	r0	0,9794	0,9796	0,982	0,9827	0,9862	0,9835	C	r0	0,9493	0,9523	0,9528	0,9507	0,9507	0,96	G	r0	0,8054	0,8303	0,8382	0,8471	0,8476	0,8586
L	f0	0,9621	0,9641	0,9649	0,9674	0,9673	0,9653	C	f0	0,9285	0,9316	0,9328	0,9297	0,9299	0,9362	G	f0	0,7885	0,8055	0,8123	0,818	0,8235	0,8283
L	p1	0,2329	0,2999	0,3298	0,3939	0,4571	0,4848	C	p1	0,3417	0,3831	0,3954	0,3743	0,3679	0,4158	G	p1	0,4637	0,5118	0,5297	0,5565	0,555	0,5692
L	r1	0,0998	0,1426	0,1416	0,1866	0,1802	0,2233	C	r1	0,2161	0,2431	0,2554	0,2379	0,2334	0,2386	G	r1	0,4145	0,4348	0,4469	0,4615	0,4741	0,4656
L	f1	0,1397	0,1933	0,1981	0,2532	0,2585	0,3057	C	f1	0,2647	0,2974	0,3103	0,2909	0,2856	0,3032	G	f1	0,4377	0,4701	0,4848	0,5046	0,5113	0,5122
A	p0	0,8273	0,8387	0,8481	0,8571	0,8556	0,8642	D	p0	0,6082	0,6587	0,6887	0,7165	0,737	0,7601	P	p0	0,6841	0,7813	0,8342	0,87	0,8795	0,8868
A	r0	0,8902	0,9016	0,9145	0,92	0,929	0,9476	D	r0	0,6328	0,6592	0,6805	0,6933	0,7029	0,7363	P	r0	0,7059	0,7821	0,8345	0,8704	0,8897	0,928
A	f0	0,8576	0,869	0,8801	0,8874	0,8908	0,904	D	f0	0,6203	0,6589	0,6845	0,7047	0,7195	0,748	P	f0	0,6948	0,7817	0,8343	0,8702	0,8846	0,9069
A	p1	0,5286	0,5776	0,6287	0,6622	0,6862	0,7546	D	p1	0,7079	0,7381	0,7576	0,7695	0,7794	0,8094	P	p1	0,8568	0,8946	0,9193	0,9375	0,9466	0,9633
A	r1	0,3986	0,4369	0,4691	0,5057	0,4973	0,5198	D	r1	0,6858	0,7377	0,7645	0,7887	0,8072	0,8281	P	r1	0,8437	0,8941	0,9192	0,9373	0,9413	0,941
A	f1	0,4545	0,4975	0,5373	0,5734	0,5767	0,6156	D	f1	0,6966	0,7379	0,7611	0,779	0,793	0,8187	P	f1	0,8502	0,8944	0,9192	0,9374	0,944	0,952

Tabla 8 SVM, precisión, sensibilidad y f1

		20	35	50	65	80	95			20	35	50	65	80	95			20	35	50	65	80	95
M	p0	0,8796	0,885	0,8767	0,8714	0,8897	0,8816	F	p0	0,9014	0,9052	0,8976	0,8982	0,9032	0,9055	Z	p0	0,6889	0,6606	0,6229	0,614	0,6528	0,6807
M	r0	0,8968	0,8295	0,7119	0,684	0,788	0,839	F	r0	0,9331	0,8781	0,7521	0,7128	0,8169	0,8428	Z	r0	0,7101	0,6727	0,6048	0,5854	0,6432	0,6755
M	f0	0,8881	0,8563	0,7858	0,7664	0,8358	0,8598	F	f0	0,917	0,8915	0,8184	0,7948	0,8579	0,873	Z	f0	0,6994	0,6666	0,6137	0,5994	0,648	0,6781
M	p1	0,3392	0,2808	0,2077	0,2023	0,2502	0,3196	F	p1	0,3513	0,2705	0,179	0,1718	0,2105	0,2413	Z	p1	0,635	0,6006	0,5403	0,527	0,5764	0,6056
M	r1	0,3014	0,3818	0,4299	0,4427	0,4199	0,4015	F	r1	0,2621	0,3296	0,3865	0,4244	0,358	0,3625	Z	r1	0,6113	0,5874	0,5592	0,5566	0,5866	0,6113
M	f1	0,3192	0,3236	0,2801	0,2777	0,3135	0,3559	F	f1	0,3002	0,2971	0,2447	0,2446	0,2651	0,2897	Z	f1	0,6229	0,5939	0,5496	0,5414	0,5814	0,6085
L	p0	0,9444	0,9477	0,9504	0,9508	0,9561	0,9616	C	p0	0,9085	0,9119	0,914	0,905	0,9083	0,922	G	p0	0,7723	0,7686	0,7539	0,7486	0,7657	0,7788
L	r0	0,9768	0,9496	0,8479	0,7995	0,899	0,9343	C	r0	0,943	0,8909	0,772	0,7288	0,8233	0,8602	G	r0	0,763	0,7009	0,6255	0,6088	0,6805	0,7167
L	f0	0,9603	0,9486	0,8962	0,8686	0,9267	0,9477	C	f0	0,9255	0,9013	0,8371	0,8074	0,8637	0,8901	G	f0	0,7677	0,7332	0,6838	0,6715	0,7206	0,7465
L	p1	0,1854	0,168	0,1047	0,0997	0,1637	0,2766	C	p1	0,3163	0,2493	0,1658	0,149	0,1819	0,217	G	p1	0,437	0,3955	0,35	0,3379	0,3846	0,4088
L	r1	0,0841	0,1627	0,2869	0,349	0,3241	0,4021	C	r1	0,2172	0,2962	0,3843	0,3829	0,3211	0,3474	G	r1	0,4498	0,4811	0,4969	0,4941	0,4895	0,4904
L	f1	0,1157	0,1653	0,1535	0,1551	0,2175	0,3277	C	f1	0,2576	0,2707	0,2317	0,2146	0,2322	0,2671	G	f1	0,4433	0,4341	0,4107	0,4014	0,4307	0,4459
A	p0	0,8314	0,8362	0,817	0,8133	0,8331	0,8364	D	p0	0,6042	0,5904	0,5408	0,5346	0,6175	0,6484	P	p0	0,6862	0,6715	0,554	0,541	0,6939	0,7813
A	r0	0,8332	0,7774	0,685	0,6527	0,7542	0,806	D	r0	0,6448	0,6564	0,6135	0,6063	0,6911	0,7193	P	r0	0,7794	0,7986	0,7298	0,7195	0,838	0,8946
A	f0	0,8323	0,8058	0,7452	0,7242	0,7917	0,8209	D	f0	0,6238	0,6216	0,5748	0,5682	0,6522	0,682	P	f0	0,7298	0,7295	0,6298	0,6176	0,7592	0,8341
A	p1	0,4494	0,4044	0,3307	0,3116	0,3841	0,4557	D	p1	0,7108	0,7102	0,6714	0,6603	0,7296	0,7695	P	p1	0,8856	0,894	0,8477	0,8408	0,9136	0,9447
A	r1	0,4463	0,4981	0,5036	0,5119	0,5036	0,5075	D	r1	0,6739	0,6491	0,6024	0,5917	0,6608	0,7062	P	r1	0,8273	0,813	0,7191	0,7082	0,8225	0,8778
A	f1	0,4478	0,4464	0,3992	0,3874	0,4358	0,4802	D	f1	0,6919	0,6783	0,635	0,6241	0,6935	0,7365	P	f1	0,8555	0,8516	0,7781	0,7688	0,8656	0,91

Tabla 9 DA, precisión, sensibilidad y f1

		20	35	50	65	80	95			20	35	50	65	80	95			20	35	50	65	80	95
M	p0	0,8769	0,8858	0,8798	0,8808	0,8763	0,8861	F	p0	0,8959	0,901	0,8982	0,8984	0,8987	0,911	Z	p0	0,6313	0,6413	0,647	0,6575	0,6441	0,6234
M	r0	0,854	0,8709	0,8787	0,8848	0,8838	0,8795	F	r0	0,875	0,8878	0,8942	0,8934	0,9017	0,9122	Z	r0	0,6177	0,6589	0,664	0,6699	0,6819	0,6612
M	f0	0,8653	0,8783	0,8792	0,8828	0,88	0,8828	F	f0	0,8854	0,8944	0,8962	0,8959	0,9002	0,9116	Z	f0	0,6244	0,65	0,6554	0,6637	0,6624	0,6417
M	p1	0,2785	0,3151	0,3212	0,3234	0,3088	0,3041	F	p1	0,2277	0,2624	0,2542	0,2672	0,2773	0,3016	Z	p1	0,5484	0,5765	0,5853	0,5815	0,5971	0,5812
M	r1	0,3196	0,346	0,3236	0,3151	0,2936	0,3177	F	r1	0,266	0,2904	0,2625	0,2779	0,2708	0,2984	Z	r1	0,5627	0,5575	0,567	0,568	0,5558	0,5407
M	f1	0,2976	0,3298	0,3224	0,3192	0,301	0,3107	F	f1	0,2454	0,2757	0,2583	0,2724	0,274	0,3	Z	f1	0,5555	0,5668	0,576	0,5747	0,5757	0,5602
L	p0	0,9448	0,9437	0,9456	0,9431	0,9423	0,9366	C	p0	0,9094	0,911	0,9148	0,9179	0,9169	0,9097	G	p0	0,7489	0,7528	0,7547	0,7531	0,753	0,7594
L	r0	0,9356	0,9307	0,9334	0,9329	0,9328	0,9205	C	r0	0,8919	0,8964	0,9084	0,9064	0,9093	0,914	G	r0	0,7167	0,727	0,739	0,7417	0,7486	0,7286
L	f0	0,9402	0,9371	0,9394	0,938	0,9375	0,9285	C	f0	0,9006	0,9037	0,9116	0,9121	0,9131	0,9119	G	f0	0,7324	0,7397	0,7468	0,7474	0,7508	0,7437
L	p1	0,1061	0,1103	0,0959	0,0899	0,0575	0,0428	C	p1	0,2282	0,2435	0,2847	0,3023	0,3232	0,2349	G	p1	0,3687	0,3768	0,3822	0,3859	0,3842	0,3513
L	r1	0,1228	0,134	0,1164	0,1053	0,0669	0,0539	C	r1	0,2644	0,2758	0,3011	0,3333	0,3446	0,2254	G	r1	0,4078	0,4087	0,4019	0,4002	0,3898	0,389
L	f1	0,1138	0,121	0,1052	0,097	0,0619	0,0477	C	f1	0,245	0,2586	0,2927	0,3171	0,3336	0,2301	G	f1	0,3873	0,3921	0,3918	0,3929	0,3869	0,3692
A	p0	0,8163	0,8219	0,8229	0,8291	0,8256	0,8198	D	p0	0,4977	0,5135	0,5243	0,521	0,5292	0,5283	P	p0	0,4595	0,4853	0,4932	0,5235	0,5572	0,5724
A	r0	0,7991	0,8168	0,8246	0,8149	0,8204	0,8078	D	r0	0,5092	0,5194	0,538	0,5379	0,5396	0,5778	P	r0	0,4885	0,4937	0,5148	0,5429	0,5565	0,5609
A	f0	0,8076	0,8194	0,8237	0,822	0,823	0,8138	D	f0	0,5034	0,5164	0,5311	0,5293	0,5344	0,552	P	f0	0,4736	0,4895	0,5038	0,533	0,5569	0,5666
A	p1	0,3894	0,4159	0,4325	0,4227	0,4247	0,3985	D	p1	0,6176	0,6261	0,6388	0,6347	0,6377	0,6712	P	p1	0,7464	0,7515	0,7627	0,779	0,7803	0,7989
A	r1	0,4159	0,4243	0,4296	0,4466	0,4334	0,4176	D	r1	0,6067	0,6205	0,6261	0,6188	0,6279	0,6255	P	r1	0,7238	0,7452	0,7468	0,7654	0,7809	0,8064
A	f1	0,4022	0,4201	0,4311	0,4344	0,429	0,4079	D	f1	0,6121	0,6233	0,6324	0,6267	0,6328	0,6475	P	f1	0,7349	0,7483	0,7547	0,7721	0,7806	0,8026

Tabla 10 Trees, precisión, sensibilidad y f1

A raíz de estos resultados queda que no hay un solo programa que sea el mejor para las distintas patologías, sino que habrá que usar el que mejor se ajuste, o la combinación que mejor se ajuste para cada una de las patologías que se van a estudiar.

Curvas ROC

Para implementar el método de las curvas ROC ya se descartaron Modic, DBE, ciática y cirugía de las clases a identificar, aquí se presentarán algunas de las curvas que se consideran mejores o más representativas para cada una de las cinco clasificaciones y después aquella que responda a la mejor puntuación, en anexos se adjuntará todo lo necesario para generar el resto de las curvas.

En primer lugar, vemos las curvas de los tres algoritmos sencillos aplicados a la artrosis teniendo como entrada sólo la geometría (las coordenadas de los puntos marcados), la leyenda se ha codificado indicando el porcentaje de entrenamiento utilizado y el algoritmo.

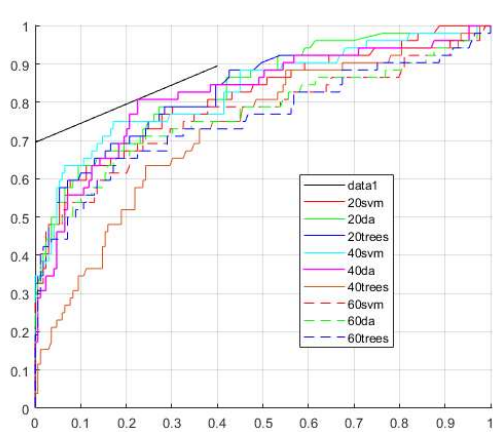


Ilustración 10 ROC geometría artrosis

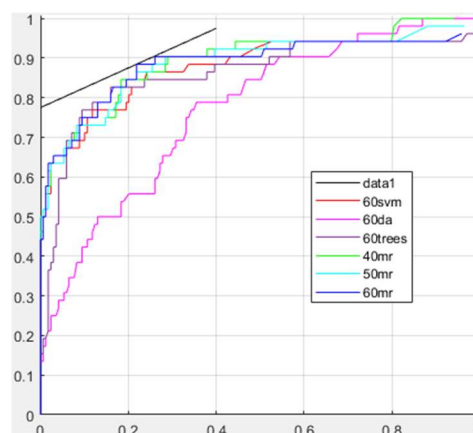


Ilustración 11 ROC artrosis

Una vez añadimos las siguientes variables como entrada: edad, peso y talla, las mejores curvas pasan a ser las del *minority report*, en concreto la del 60% de entrenamiento. Se puede ver una mejora significativa, del 80% de TPR y 22% FPR pasamos al 88,5% de FPR y 20% de FPR. Continuando con la Listesis, los mejores resultados son los siguientes, a la izquierda (ilustración 11) sólo con la geometría, a la derecha (ilustración 12) con las demás variables

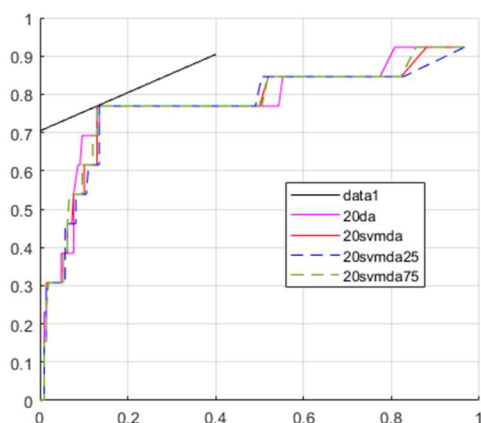


Ilustración 12 ROC geometría listesis

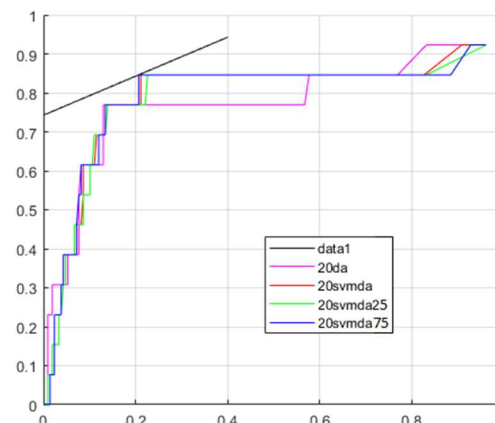


Ilustración 13 ROC listesis

Con la introducción de las nuevas variables se consigue una mejora del TPR de 7.7% a costa de un aumento en el FPR de también un 7.7% pero al valorar más el TPR es considerada mejor. Análogamente para la estenosis C, se muestran los resultados en las ilustraciones 13 y 14.

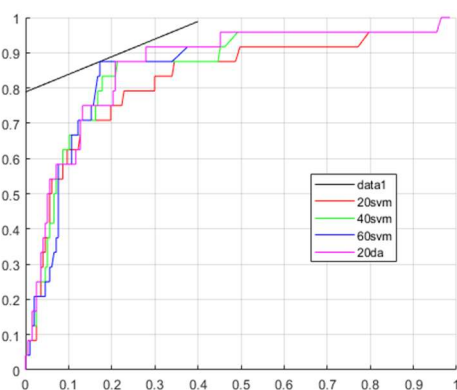


Ilustración 14 ROC geometría estenosis C

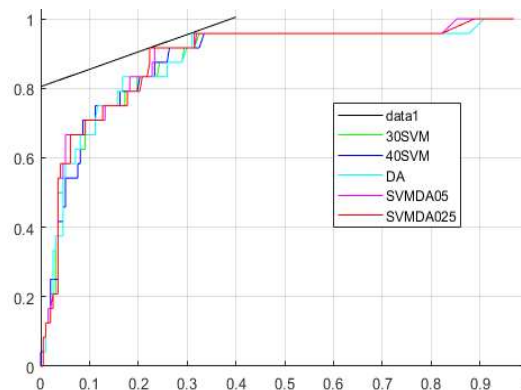


Ilustración 15 ROC estenosis C

El porcentaje de entrenamiento para la curva de la ilustración 15 es de 20% si no se especifica lo contrario. El punto elegido en la curva de la geometría corresponde a un FPR de 0.173 y un TPR de 0.875 mientras que con las nuevas entradas cambia a 0.22 y 0.917, respectivamente.

Para el caso de la Estenosis F tenemos:

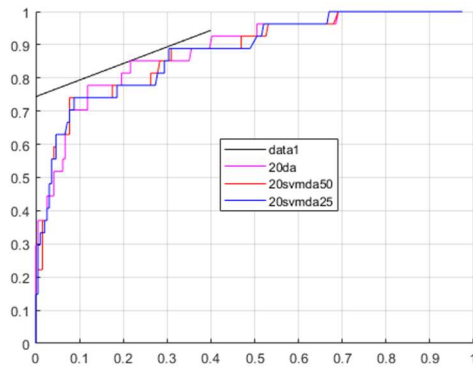


Ilustración 16 ROC geometría estenosis F

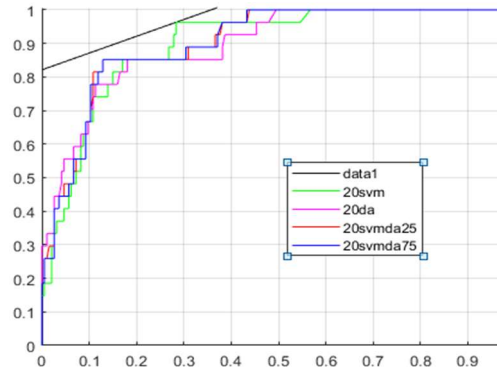


Ilustración 17 ROC estenosis F

Donde se obtiene un mejor punto de 0.22-0.84 en geometría (ilustración 15) y 0.28-0.96 en la completa (ilustración 16).

Por último, pero no por ello menos importante se presentan los datos de patología. Cabe mencionar que el resultado en la curva que tiene en cuenta al DBE y la ciática se podría decir que es demasiado bueno, porque en el conjunto de casos solo hay un paciente que tiene DBE y no es patológico y uno con ciática que no es patológico, sin embargo, los resultados al mirar únicamente la geometría son también muy buenos, aproximándose mucho a la curva ideal, con un FPR de 0.042 y un TPR de 0.96

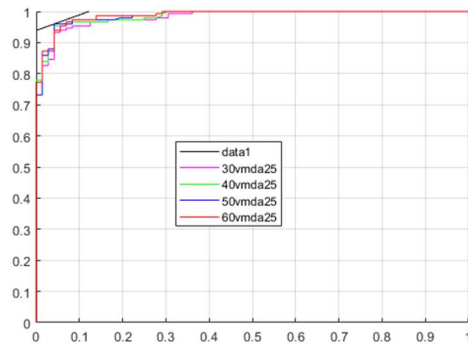


Ilustración 18 ROC geometría patología

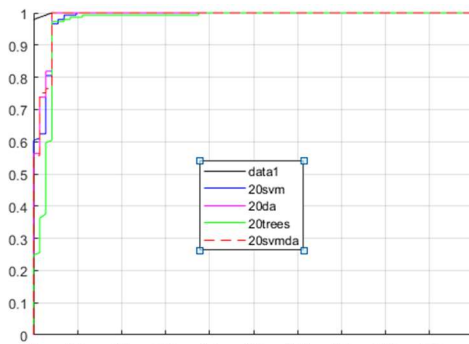
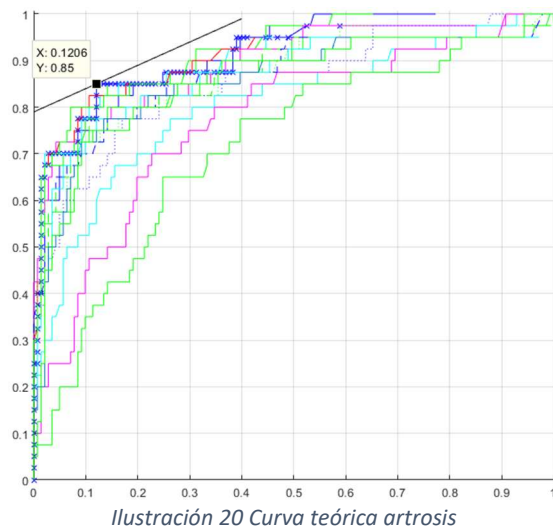


Ilustración 19 ROC patología

Verificación

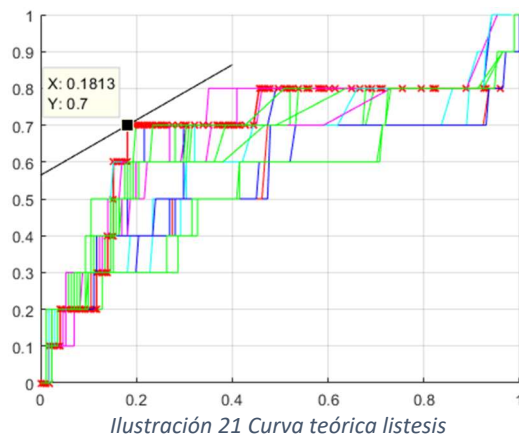
A continuación, se van a presentar la mejor curva obtenida con el grupo de modelado y la tabla que contiene el FPR, TPR y umbral del mejor punto teórico, junto con la matriz de confusión (recordar ilustración 1 y ecuaciones 2 y 4) obtenida tras ejecutar el programa 3.000 veces. Los umbrales se redondearán ligeramente a la baja para tener cierta seguridad de que detectamos a los máximos enfermos posibles.



FPR teórico	TPR teórico	Umbral
0.1206	0.85	0.1805

20	8
2	10

Como se puede comprobar el TPR es de 0.83 y el FPR de 0.28, no son unos resultados tan buenos como los dados por la curva, pero se aproximan y son de por sí buenos. Se ha utilizado un umbral de 0.18 y el método: SVM con grupos formados por el 40% del conjunto de entrenamiento.



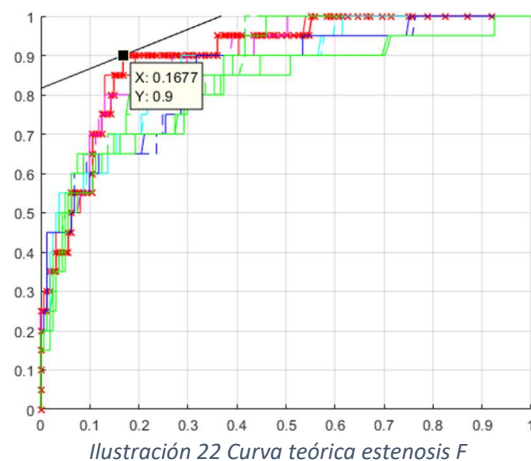
0.1813	0.7	0.0565
--------	-----	--------

28	9
0	3

Con un umbral de 0.055 obtenemos un 100% de detección y 0.19 de FPR.

Para este caso se ha ajustado el umbral a 0.75 porque si no el punto elegido se consideraba con un FPR demasiado alto. El método utilizado fue una media ponderada entre SVM y DA a favor de DA 2-1.

Cabe decir que, muy destacado sobre todo en las dos curvas recientemente presentadas, el hecho de retirar del conjunto de modelado a 40 pacientes (18% del total) para tener un grupo de verificación hace que los modelos se vean perjudicados, al comparar estas curvas con las del punto anterior, prueba de que el conjunto de pacientes con el que trabajamos es escaso y los resultados podrían mejorar significativamente de aumentar este.



0.168	0.900	0.0745
-------	-------	--------

29	4
1	6

Ajustando el umbral a 0.074 se obtienen un TPR de 0.833 y un FPR de 0.121.

El método elegido fue la media entre SVM y DA, con 20% de entrenamiento.

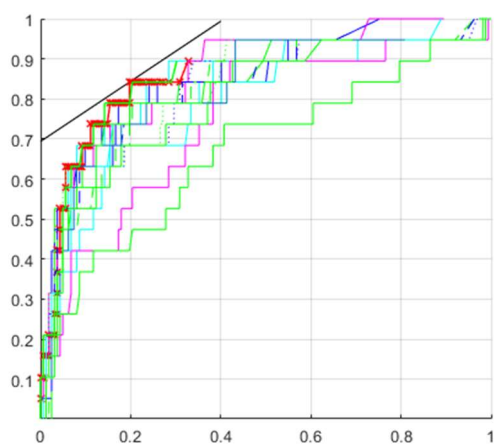


Ilustración 23 Curva teórica estenosis C

0.1975	0.8421	0.032
--------	--------	-------

30	5
1	4

Umbral 0.03. TPR 0.8, FPR 0.143.

Para este caso se ha ajustado la pendiente de la recta que determina el mejor punto a 0.75 porque si no el punto elegido se consideraba con un FPR demasiado alto.

El método usado ha sido SVM con 60% de entrenamiento.

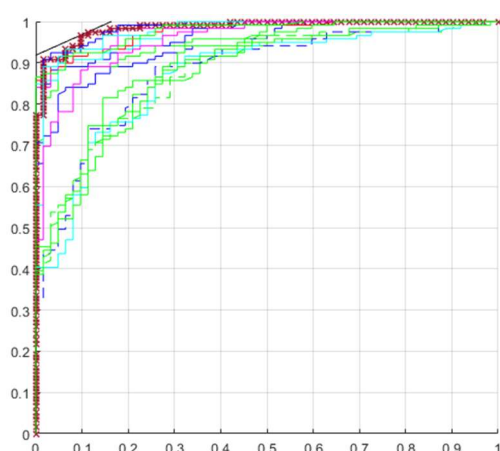


Ilustración 24 Curva teórica patología

0.113	0.975	0.428
-------	-------	-------

6	4
1	29

En este caso se puede apreciar una anomalía que será estudiada en un futuro y es el FPR tan alto que presenta, 0.4 con respecto al teórico, en cuanto al TPR nos encontramos con una cifra de 0.967. El umbral utilizado ha sido 0.4. Se ha usado DA con un 30% de entrenamiento.

Hay que tener en cuenta que estos estudios se han realizado con pocos datos y en algunos casos, que se diagnostique correctamente un paciente más o un paciente menos conlleva un cambio relativamente grande, como puede pasar con los pacientes sanos de la patología, o los listésicos enfermos en los que tenemos una sensibilidad del 100%. Se ha comprobado que los diagnósticos tienen repetibilidad, con un mismo umbral y método se tienen resultados iguales.

Conclusiones

En conclusión, se puede destacar que los resultados son muy satisfactorios, con detecciones que varían entre el 80 y 100% sin haber sacrificado a penas la sensibilidad que oscila entre 75 y 85 % para todas las clasificaciones. Este método de repetición de los modelados para la obtención de probabilidades en vez de clasificaciones binarias se ha probado muy útil en conjunto con el uso de las curvas ROC para la elección del umbral. Es un método que ha sido desarrollado expresamente para este proyecto, y en el cual me gustaría profundizar, ya que al no haber encontrado más información sobre él creo que tiene aún mucho potencial y se podría aplicar a proyectos con problemáticas similares, aunque no compartan el área de aplicación. Asimismo, con este trabajo he aprendido mucho, gracias a todos los errores que he ido cometiendo y las soluciones que he ideado para estos, se hayan probado útiles o no.

Para dar una explicación a los fallos en la clasificación de los pacientes sanos en el grupo patología se ha estudiado la naturaleza de estos cuatro pacientes y realmente son pacientes excepcionales, uno de ellos estaba marcado como afectado en DBE pero se consideraba sano, de los cuales solo hay dos casos en el total de pacientes, mientras que los hay que con DBE son considerados patológicos, otros dos entran en el grupo de riesgo de pacientes que sufren de obesidad y el paciente restante también pertenece a otro grupo de riesgo que es el de los pacientes de avanzada edad. Este tipo de incidencias tendrían menos importancia relativa si se tuviese una base de datos más grande, otra de las razones para visitar el proyecto una vez se hayan adquirido más datos.

Futuro

A partir de aquí hay varias rutas que se pueden tomar, por un lado, siempre es posible afinar más el programa, encontrar cuáles son los datos más significativos, quizás introducir la talla, o el IMC en el conjunto de entrada. Por otro lado, aumentar el número de pacientes con los que construir los modelos, lo cual ahora es difícil porque los marcadores fueron realizados a mano por el médico y eso conlleva mucho tiempo, pero cuando se consiga automatizar ese proceso veríamos un gran aumento en los datos con los que trabajar. También el desarrollo de una red neuronal para ser usada junto a los demás algoritmos podría ayudar a conseguir resultados mejores.

Sin embargo, el ajuste fino o mejora de este proyecto en este marco no es lo único que cabe considerar. Aunque muchas de las soluciones que se han desarrollado para los problemas presentados en este trabajo ya existieran, el método de utilizar diferentes subgrupos del total de elementos de entrenamiento, en vez de crear un modelo que utilice todos, para así tener muchos modelos distintos con los que poder generar las probabilidades de pertenencia a la clase a estudiar es algo de lo que no he encontrado evidencia de uso, es parecido al *Boosting* (2) y al *Bootstrapping* (1) pero no creo que ninguna de estas técnicas sea exactamente lo que se ha hecho, o al menos no la forma convencional de aplicarlas. Puede que se pruebe como una solución para las ocasiones en las que los elementos de entrenamiento sean escasos y haya poco balance entre las clases.

Bibliografía

1. **Tejada, Dr. Serrano Sáenz de.** espaldaycuello.com. [En línea] 6 de Julio de 2015. [Citado el: 17 de Junio de 2019.] <https://espaldaycuello.com/listesis/>.
2. **mayo clinic.** www.mayoclinic.org. [En línea] 21 de Septiembre de 2018. [Citado el: 17 de Junio de 2019.] <https://www.mayoclinic.org/es-es/diseases-conditions/spinal-stenosis/symptoms-causes/syc-20352961>.
3. **physicianpartnersofamerica.** www.physicianpartnersofamerica.com. [En línea] [Citado el: 20 de Junio de 2019.] <https://www.physicianpartnersofamerica.com/es/medical-services/laser-spine-surgery/conditions/foraminal-stenosis/>.
4. **Paradells, Dr. Víctor Rodrigo.** www.cun.es. [En línea] [Citado el: 20 de Junio de 2019.] <https://www.cun.es/enfermedades-tratamientos/enfermedades/estenosis-canal>.
5. **Red Española de Investigadores de Espalda.** www.espalda.org. [En línea] 29 de Marzo de 2019. [Citado el: 17 de Junio de 2019.] <http://www.espalda.org/divulgativa/dolor/causas/alteraciones/vertebral.asp>.
6. **mayo clinic.** www.mayoclinic.org. [En línea] 30 de Mayo de 2018. [Citado el: 17 de Junio de 2019.] <https://www.mayoclinic.org/es-es/diseases-conditions/sciatica/symptoms-causes/syc-20377435>.
7. **medtronica iberia.** medtronic.com. [En línea] 30 de Mayo de 2017. [Citado el: 17 de Junio de 2019.] <https://www.medtronic.com/es-es/tu-salud/patologias/discopatia-degenerativa-lumbar.html>.
8. **Espalda, Red Española de Investigadores de Dolencias de.** espalda.org. [En línea] 27 de Junio de 2016. [Citado el: 17 de Junio de 2019.] <http://www.espalda.org/divulgativa/dolor/causas/alteraciones/signosmodic.asp>.
9. **mathworks.com.** [En línea] [Citado el: 1 de Julio de 2019.] <https://es.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html?lang=en>.
10. **Matlab.** youtube. [En línea] 1 de Octubre de 2018. [Citado el: 1 de Julio de 2019.] <https://www.youtube.com/watch?v=eBm8Uo9yhwI>.
11. **Mathworks.** mathworks.com. [En línea] [Citado el: 18 de Julio de 2019.] <https://es.mathworks.com/help/stats/fitctree.html>.
12. **mathworks.** mathworks.com. [En línea] [Citado el: 4 de Julio de 2019.] <https://es.mathworks.com/help/stats/confusionmat.html>.
13. **rocca, baptiste and joseph.** towardsdatascience. [En línea] 27 de Enero de 2019. [Citado el: 8 de Julio de 2019.] <https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28>.
14. **Narkhede, Sarang.** towardsdatascience. [En línea] 26 de Junio de 2018. [Citado el: 15 de Julio de 2019.] <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>.
15. **NCSS.** *NCSS statistical software Chapter 546.*